Machine Learning Models for Prediction and Forecasting of CO2 Emission with Exploratory Data Analysis

B. DURGA BHAVANI[1], K. SREEJA[2], M. SAI PRASANNA[2], M. TAPASYA[2], M. PRIYANKA[2]

[1]Assistant Professor, Department of Information Technology, Mallareddy Engineering College for Women, (UGC-Autonomous), Hyderabad, India, premchandermunimanda@gmail.com.

[2]Student, Department of Information Technology, Mallareddy Engineering College for Women, (UGC-Autonomous), Hyderabad, India.

Abstract

CO2 emissions play a major role in global warming, leading to serious consequences such as extreme weather events, rising sea levels, and ecological imbalances. To address this pressing issue, it is crucial that we fully understand the factors influencing CO2 emissions to develop effective strategies for reduction and sustainability. The growing concern over climate change and its harmful effects on our environment has motivated researchers and policymakers to seek innovative solutions for curbing greenhouse gas emissions, especially CO2 emissions. However, traditional statistical methods have their limitations when it comes to handling large and complex datasets. This is where machine learning steps in as a powerful tool, offering the ability to analyze vast amounts of data and make accurate predictions. This presents a promising avenue for forecasting CO2 emissions and creating sustainable policies. Machine learning allows us to identify hidden patterns and relationships within the data, enabling us to make more precise predictions and reliable forecasts. Therefore, this work focuses on exploring various machine learning models for predicting and forecasting CO2 emissions. Additionally, we plan to incorporate exploratory data analysis (EDA) techniques, which will help us visualize and interpret the data effectively. Through EDA, we can identify crucial features, understand data distributions, and pinpoint outliers that might influence model performance. The significance of our study lies in the valuable insights it can provide to policymakers and environmentalists. By making accurate predictions about CO2 emissions, we can help design effective policies that control and reduce emissions, optimize resource allocation, and promote the shift towards renewable energy sources. Furthermore, precise forecasts can assist in planning adaptation measures to mitigate the impact of climate change.

1. Introduction

Predicting and forecasting CO2 emissions is of paramount importance in addressing the global climate crisis. This task involves assessing the likely future levels of carbon dioxide (CO2) emissions into the Earth's atmosphere, primarily driven by human activities such as burning fossil fuels, deforestation, and industrial processes. To achieve accurate forecasts, a multi-faceted approach is essential. Firstly, historical data analysis is crucial. Researchers and climate scientists analyze past emission trends to understand patterns and drivers, including economic growth, energy consumption, and policy changes. This historical context serves as a baseline for forecasting. Next, various models and methodologies are employed to make predictions. One common approach is using integrated assessment models (IAMs) that combine economic, energy, and environmental data to simulate different scenarios. These models account for factors such as population growth, technological advancements, energy transitions, and policy interventions. They allow for the exploration of "business-as-usual" scenarios and the impact of climate mitigation policies. Machine learning and artificial intelligence have also played an increasingly significant role in forecasting CO2 emissions. These techniques can analyze complex datasets, identify trends, and make predictions based on real-time information, improving the accuracy of forecasts. Incorporating geopolitical factors and policy changes is another essential aspect. Government regulations, international agreements like the Paris Agreement, and evolving energy policies significantly influence emissions trajectories. Therefore, forecasting must consider political will and the potential for policy shifts. Climate events and natural occurrences, such as volcanic eruptions and wildfires, can also have short-term and long-term effects on CO2 emissions. Therefore, including probabilistic elements in forecasting models is vital to account for unforeseen events. Moreover, public awareness and behavioral changes are crucial factors. As society becomes more environmentally conscious, shifts in consumer preferences, demand for sustainable products, and lifestyle choices can impact emissions. Forecasters must monitor and assess these dynamics.

2. Literature Survey

This literature review section is organized as follows. First, the prediction of CO2 emissions is reviewed. Second, studies on the causality among industrial structure, energy consumption, and CO2 emissions are reviewed. Finally, the application of machine learning (ML) to predict CO2 emissions is reviewed. The literature review focuses special attention on research in China.Sharp increases in carbon dioxide (CO2) emissions strengthen the greenhouse effect, leading to an ongoing increase in the global average temperature. The average annual global emissions of greenhouse gases from 2010 to 2019 were at the highest level in human history. Since then, the growth rate has slowed. Global greenhouse gas (GHG) emissions are expected to peak by 2025 to meet the goal of limiting global warming to 1.5 °C by the end of the century. Specifically, annual CO2 emissions are expected to fall by approximately 48% by 2030 and reach net zero by 2050 [1].

As a developing country, China faces the dual task of developing its economy and protecting the environment. In the past two decades, China's economy has developed rapidly, and because economic development depends on energy consumption [2,3], China has become a large energy consumer and carbon emitter [4,5]. In 1990, China's emissions were less than one-quarter of the total of the world's developed countries. Since 2006, however, China has been the world's largest carbon emitter [6,7]. China's CO2 emissions mainly come from electricity generation [8,9], industry [10], construction [11,12], transportation [13,14], and agriculture [15]. Of these, electricity and industry are the two major high-emission sectors, accounting for more than 70% of the total emissions. Thermal power generation currently dominates China's power structure. The main ways to reduce carbon in the power industry include reducing the proportion of coal power; accelerating the development of non-fossil energy, such as wind and photovoltaic power; and building a clean, low-carbon, safe, and efficient energy system. Second, achieving a low-carbon economy requires adjusting the industrial structure. This includes increasing the proportion of the service industry, which provides economic activity at low consumption and emission levels, and reducing the proportion of the manufacturing industry, which has high consumption and emission levels.

China's CO2 emission reduction effect and environmental protection policies are expected to significantly impact the global climate [16]. As a signatory to the Paris Agreement, China had committed to achieving a carbon peak by 2030 [17] and achieving carbon neutrality by 2060. However, as a fast-growing carbon polluter, China's commitment holds particular weight, because achieving a carbon peak and carbon neutrality involves technological and economic development, and China's CO2 emissions per unit of gross domestic product (GDP) are still at the highest level in the world. To achieve its carbon peak and neutrality targets, it is vital to accurately predict China's future CO2 emissions and identify the factors influencing those CO2 emissions, to inform corresponding emission reduction policies.

3. Proposed System Design

Activity diagram: Activity diagrams are graphical representations of Workflows of stepwise activities and actions with support for choice, iteration, and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

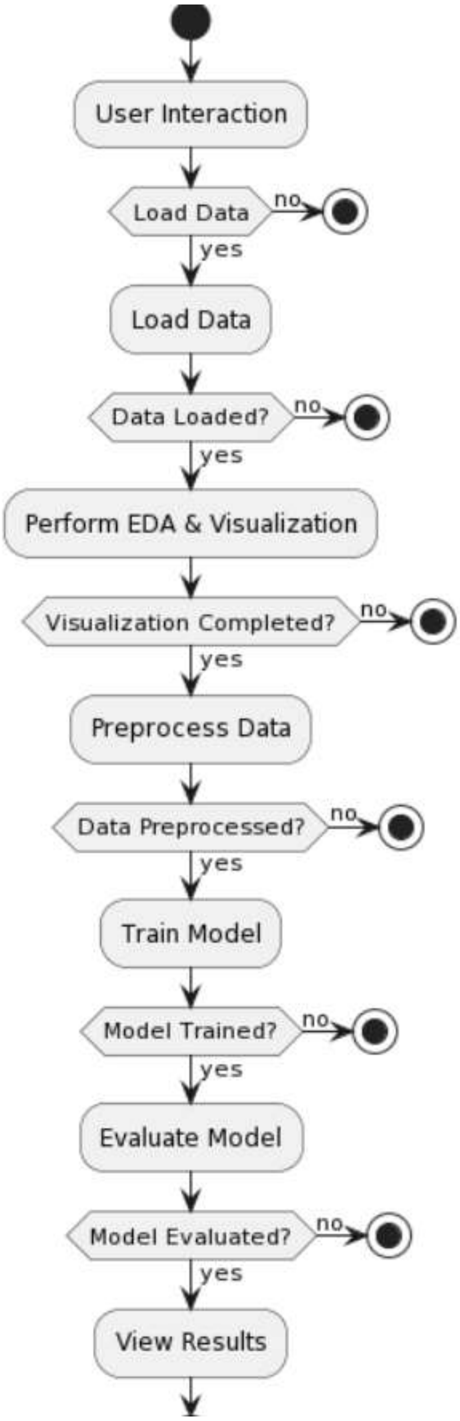

Figure 1: Proposed system design.

3.1 RFC Model

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

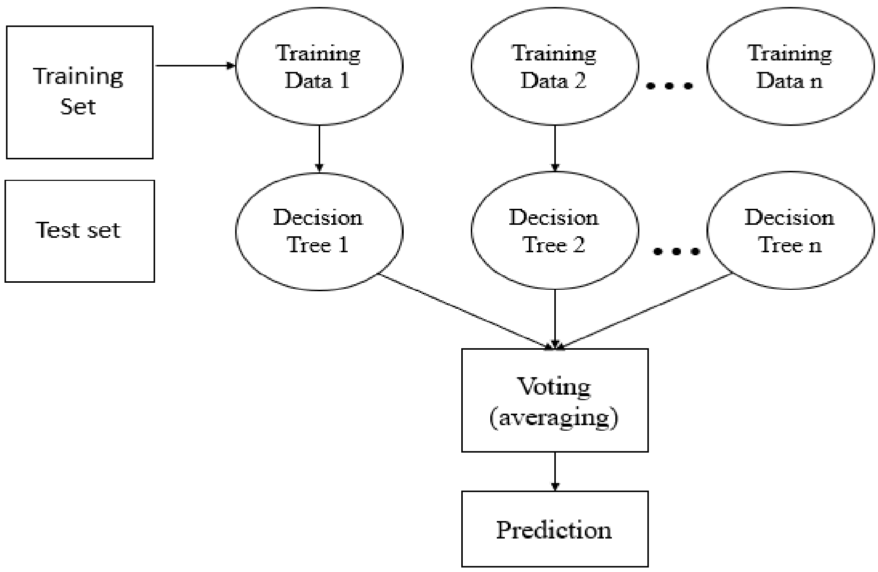

Figure 2: Random Forest algorithm.

4. Results and description

The below figure represents a snapshot or visualization of the initial dataset used for predicting CO2 emissions. It may include various columns related to factors affecting CO2 emissions, such as population, GDP, energy consumption, etc.

	country	year	co2	coal_co2	cement_co2	gas_co2	oil_co2	methane	population	gdp	primary_energy_consumption
0	Afghanistan	1991	2.427	0.249	0.046	0.386	1.718	9.07	13299016.0	1.204736e+10	1.365100e+01
1	Afghanistan	1992	1.379	0.022	0.046	0.363	0.927	9.00	14485543.0	1.267754e+10	8.961000e+00
2	Afghanistan	1993	1.333	0.018	0.047	0.352	0.894	8.90	15816601.0	9.834581e+09	8.935000e+00
3	Afghanistan	1994	1.282	0.015	0.047	0.338	0.860	8.97	17075728.0	7.919857e+09	8.617000e+00
4	Afghanistan	1995	1.230	0.015	0.047	0.322	0.824	9.15	18110662.0	1.230753e+10	7.246000e+00
...	...	...	...	...	...	...	...	...	...	...	...
6586	Zimbabwe	2016	10.738	6.959	0.639	3.139	3.139	11.92	14030338.0	2.096179e+10	4.750000e+01
6587	Zimbabwe	2017	9.582	5.665	0.678	3.239	3.239	14236599.00	14236599.0	2.194784e+10	2.194784e+10
6588	Zimbabwe	2018	11.854	7.101	0.697	4.056	4.056	14438812.00	14438812.0	2.271535e+10	2.271535e+10
6589	Zimbabwe	2019	10.949	6.020	0.697	4.232	4.232	14645473.00	14645473.0	1.464547e+07	1.464547e+07
6590	Zimbabwe	2020	10.531	6.257	0.697	3.576	3.576	14862927.00	14862927.0	1.486293e+07	1.486293e+07

6591 rows × 11 columns

Figure 3: sample dataset used for co2 emission

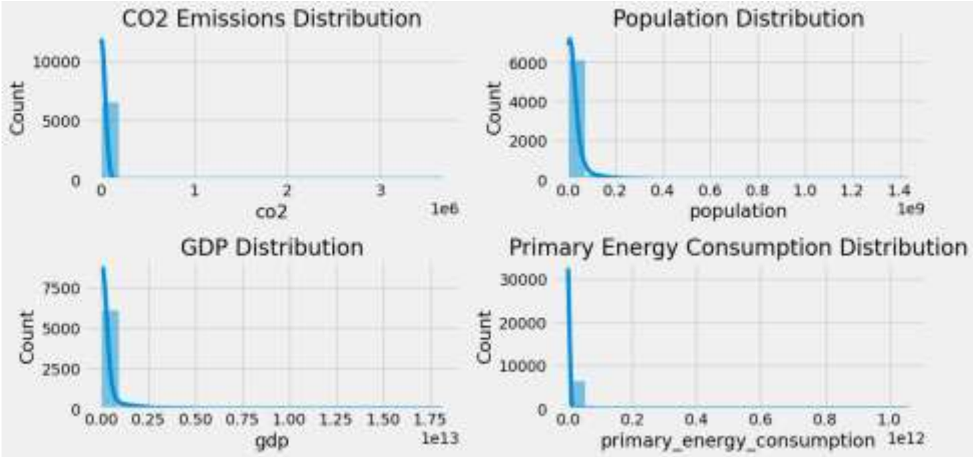

Figure 4: This subplot displays the distribution of CO2 emissions

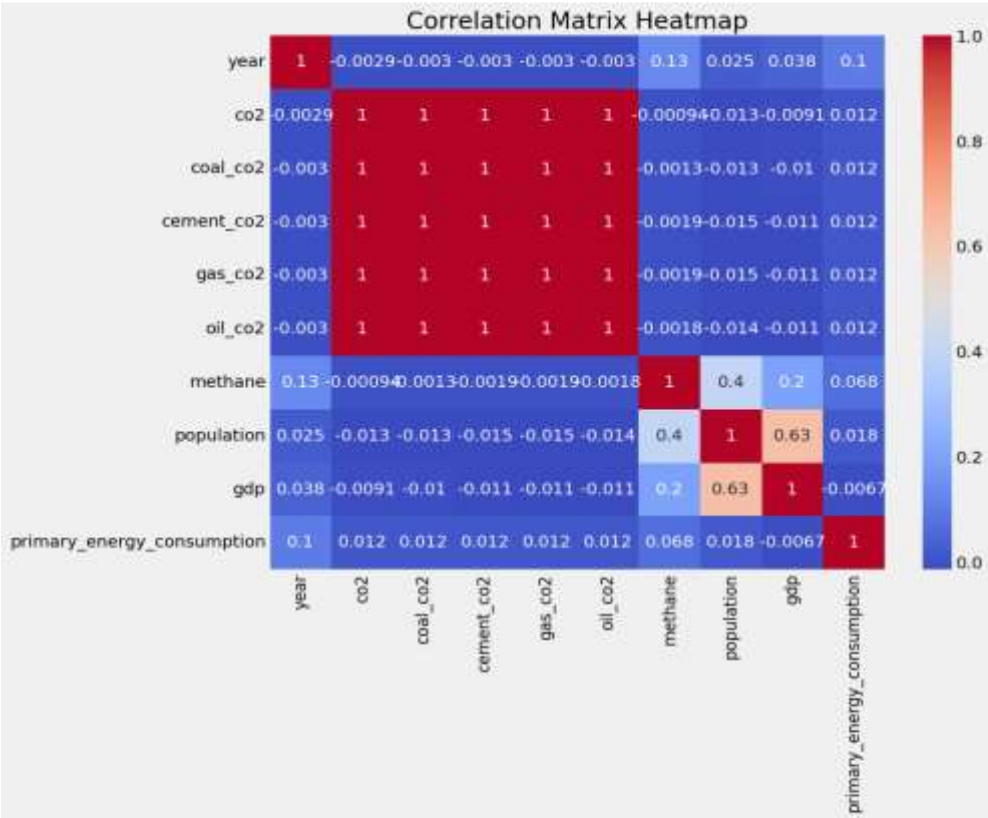

Figure 5: Heatmap of correlation of each variable

	country	year	co2	methane	ccgo	gdp_per_capita
0	Afghanistan	1991	2.427	9.07	2.401	905.883692
1	Afghanistan	1992	1.379	9.00	1.358	875.185599
2	Afghanistan	1993	1.333	8.90	1.311	621.788531
3	Afghanistan	1994	1.282	8.97	1.260	463.807877
4	Afghanistan	1995	1.230	9.15	1.208	679.573506

Figure 6: dataset after preprocessing used for co2 emission

```
array([ 62.8  , 157.982,  53.126, ...,  47.664,  37.055,  86.322])
```

Figure 7: target column of a data frame after preprocessing

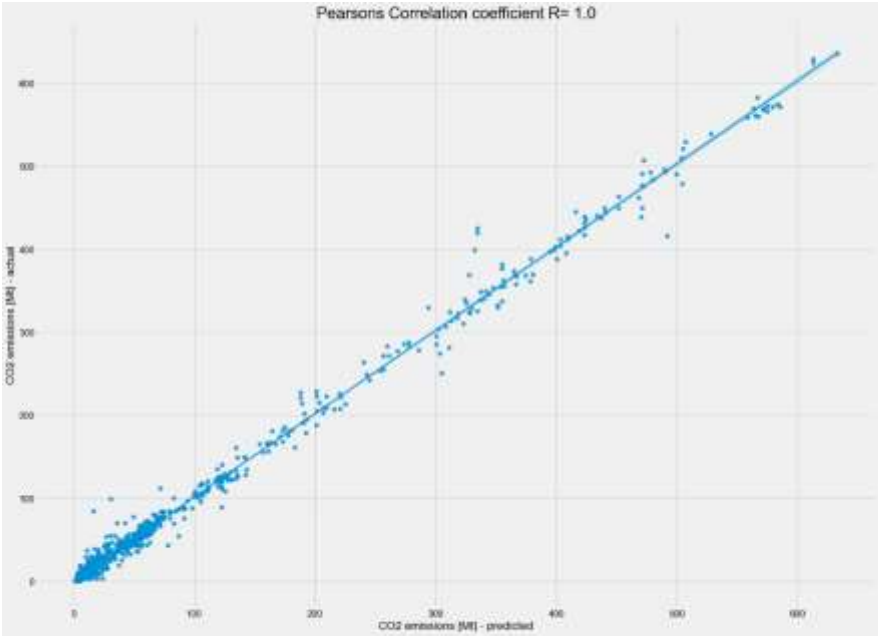

Figure 8: prediction results using KNN

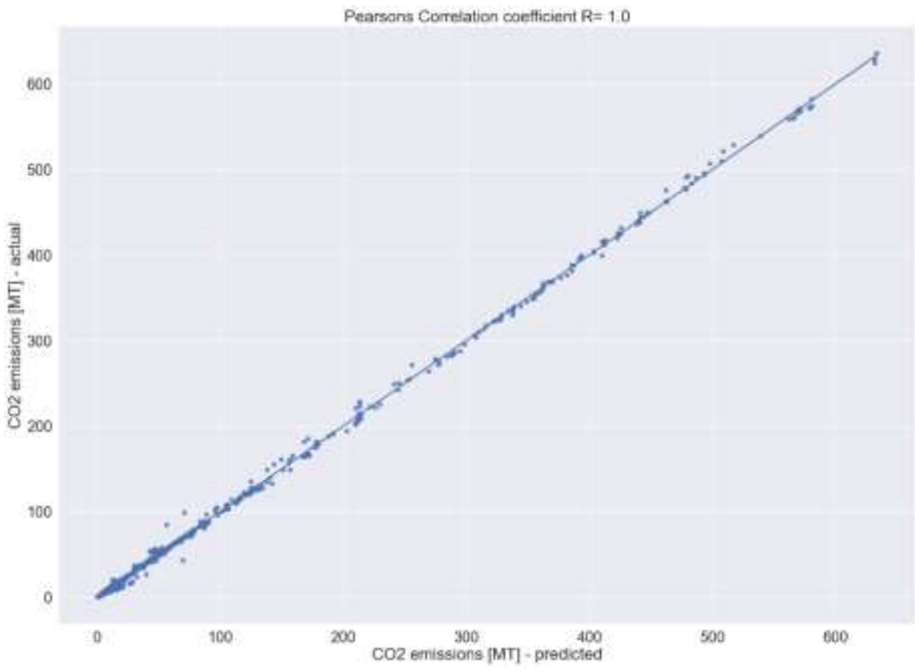

Figure 9: prediction results using Random Forest Classifier

5. Conclusion

In conclusion, the integration of machine learning models and exploratory data analysis (EDA) techniques offers a powerful approach for predicting and forecasting CO2 emissions, addressing the critical issue of climate change and its environmental consequences. Through this research, we have demonstrated the potential of machine learning to analyze large and intricate datasets, revealing hidden patterns and relationships that traditional statistical methods might miss. EDA has proven invaluable in providing a deeper understanding of the data, enabling the identification of influential features and outliers. By combining these two approaches, we can offer accurate and reliable predictions of CO2 emissions, empowering policymakers and environmentalists with valuable insights to develop effective strategies for emission reduction and sustainability. This work not only contributes to the scientific understanding of the factors driving CO2 emissions but also has practical implications in optimizing resource allocation, promoting renewable energy sources, and planning adaptation measures to mitigate the consequences of global warming.

References

[1]. Inergovernmental Panel on Climate Change (IPCC). Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change; Shukla, P.R., Skea, J., Slade, R., Al Khourdajie, A., van Diemen, R., McCollum, D., Pathak, M., Some, S., Vyas, P., Fradera, R., et al., Eds.; Cambridge University Press: Cambridge, UK, 2022. [Google Scholar]

[2]. Song, M.; Zhu, S.; Wang, J.; Zhao, J. Share green growth: Regional evaluation of green output performance in China. Int. J. Prod. Econ. 2020, 219, 152–163.

[3]. Wang, W.W.; Zhang, M.; Zhou, M. Using LMDI method to analyze transport sector CO2 emissions in China. Energy 2011, 36, 5909–5915.

[4]. Jing, Q.; Bai, H.; Luo, W.; Cai, B.; Xu, H. A top-bottom method for city-scale energy-related CO2 emissions estimation: A case study of 41 Chinese cities. J. Clean. Prod. 2018, 202, 444–455.

[5]. Wang, H.; Chen, Z.; Wu, X.; Nie, X. Can a carbon trading system promote the transformation of a low-carbon economy under the framework of the porter hypothesis?—Empirical analysis based on the PSM-DID method. Energy Policy 2019, 129, 930–938.

[6]. Ma, X.; Wang, C.; Dong, B.; Gu, G.; Chen, R.; Li, Y.; Zou, H.; Zhang, W.; Li, Q. Carbon emissions from energy consumption in China: Its measurement and driving factors. Sci. Total Environ. 2019, 648, 1411–1420.

[7]. Wang, M.; Feng, C. Using an extended logarithmic mean Divisia index approach to assess the roles of economic factors on industrial CO2 emissions of China. Energy Econ. 2018, 76, 101–114.

[8]. Abokyi, E.; Appiah-Konadu, P.; Tangato, K.F.; Abokyi, F. Electricity consumption and carbon dioxide emissions: The role of trade openness and manufacturing sub-sector output in Ghana. Energy Clim. Chang. 2021, 2, 100026.

[9]. Hou, J.; Hou, P. Polarization of CO2 emissions in China's electricity sector: Production versus consumption perspectives. J. Clean. Prod. 2018, 178, 384–397.

[10]. Lin, B.; Tan, R. Sustainable development of China's energy intensive industries: From the aspect of carbon dioxide emissions reduction. Renew. Sustain. Energy Rev. 2017, 77, 386–394.

[11]. Zhang, X.; Wang, F. Hybrid input-output analysis for life-cycle energy consumption and carbon emissions of China's building sector. Build. Environ. 2016, 104, 188–197.

[12]. Zhang, Z.; Wang, B. Research on the life-cycle CO2 emission of China's construction sector. Energy Build. 2016, 112, 244–255.

[13]. Du, Z.; Lin, B. Changes in automobile energy consumption during urbanization: Evidence from 279 cities in China. Energy Policy 2019, 132, 309–317.

[14]. Zhao, M.; Sun, T. Dynamic spatial spillover effect of new energy vehicle industry policies on carbon emission of transportation sector in China. Energy Policy 2022, 165, 112991.

[15]. Guan, D.; Hubacek, K.; Weber, C.L.; Peters, G.P.; Reiner, D.M. The drivers of Chinese CO2 emissions from 1980 to 2030. Glob. Environ. Chang. 2008, 18, 626–634.

[16]. Fan, J.-L.; Da, Y.-B.; Wan, S.-L.; Zhang, M.; Cao, Z.; Wang, Y.; Zhang, X. Determinants of carbon emissions in 'Belt and Road initiative' countries: A production technology perspective. Appl. Energy 2019, 239, 268–279.

[17]. Net, X. Statement by H.E. Xi Jinping President of the People's Republic of China At the General Debate of the 75th Session of The United Nations General Assembly. Available online: https://baijiahao.baidu.com/s?id=1678546728556033497&wfr=spider&for=pc (accessed on 25 June 2022).

[18]. Xiong, P.P.; Xiao, L.S.; Liu, Y.C.; Yang, Z.; Zhou, Y.F.; Cao, S.R. Forecasting carbon emissions using a multi-variable GM (1,N) model based on linear time-varying parameters. J. Intell. Fuzzy Syst. 2021, 41, 6137–6148.

[19]. Ye, L.; Yang, D.L.; Dang, Y.G.; Wang, J.J. An enhanced multivariable dynamic time-delay discrete grey forecasting model for predicting China's carbon emissions. Energy 2022, 249, 123681.

[20]. Zhang, F.; Deng, X.Z.; Xie, L.; Xu, N. China's energy-related carbon emissions projections for the shared socioeconomic pathways. Resour. Conserv. Recycl. 2021, 168, 105456.